Правительство Российской Федерации
Федеральное государственное автономное образовательное
учреждение высшего профессионального образования
Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»
Департамент анализа данных и искусственного интеллекта

КУРСОВАЯ РАБОТА

на тему

Рекомендательные системы на основе доверия

Выполнил студент группы БПМИ133
Стеценко Макар Александрович

Научный руководитель:
Доцент, кандидат технических наук,
Игнатов Дмитрий Игоревич

Москва 2016

Оглавление

Аннотация

"Исследование возможности встраивания контекстной информации в алгоритмы коллаборативной фильтрации на основе матричных разложений"

Стеценко Макар Александрович
Факультет компьютерных наук. Второй курс. 203 группа. 2015 год.
Игнатов Дмитрий Игоревич
Факультет компьютерных наук. Департамент анализа данных и искусственного интеллекта. Научный руководитель. dignatov@hse.ru

Наличие контекстной информации является одним из важнейших факторов для построения личных рекомендаций. Однако, классические алгоритмы коллаборативной фильтрации, основанные на матричных разложениях, таких как SVD разложение, используют только информацию о пользователях и предметах и не предоставляют явных методов включения дополнительных факторов. В данной работе будет показан один из методов встраивания контекстной информации в алгоритм, использующий SVD разложение. Для тестирования рассматриваемого метода будет использоваться открытый банк данных MovieLens. База данных содержит пользователей портала MovieLens, каждый из которых оценил не менее 20 фильмов, а так же информацию о каждом фильме.

Abstract

"Integrating Contextual Information into Collaborative Filtering Algorithms based on Matrix Decomposition"

Stetsenko Makar
Faculty of Computer Science. 2nd course. 203 group. 2015 year.
Ignatov Dmitry
Faculty of Computer Science. School of Data Analysis and Artificial Intelligence.
Scientific adviser. dignatov@hse.ru

Context has always been an important factor in personalized Recommender systems. However, standard collaborative filtering algorithms based on matrix factorization rely mainly on user and subject information and don't provide any methods for incapsulating extra data. This work demonstrates such a method based on SVD decomposition. To test results an open data base taken from MovieLens is used. The database provides information about users and movies.

Введение

Рекомендательные системы стали неотъемлемой частью жизни людей и бизнеса, особенно после популяризации BigData и открытых данных. Задача таких систем угадывать предпочтения пользователей основываясь на их предыдущих действиях. Более формально, задача о рекомендациях состоит в угадывании значения пары (пользователь, предмет). В простейшем варианте, ответом на данный вопрос будет либо 1 - предмет интересен пользователю, либо 0 - предмет не стоит рекомендовать.

Коллаборативная фильтрация один из способов построения рекомендаций. Данный метод использует информацию о других пользователях системы, чтобы понять, какой предмет скорее всего понравится рассматриваемому пользователю. Несмотря на то, что упрощенная модель *(пользователь, предмет)* подходит для моделирования многих ситуаций, очень часто приходится сталкиваться с дополнительными параметрами, которые играют важную роль при решении задачи о рекомендациях. Таким параметром может быть время, тогда уже имеется 3-х мерный вектор *(пользователь, предмет, время)*. Множество переменных, которые влияют на отношение пользователя к предмету и следовательно на рекомендации для этого пользователя, называется *контекстом*.

Целью данной работы будет решение задачи о рекомендациях с использованием контекстной информации.

Глава 1

1.1 Введение

Рекомандательные системы стали назависимой областью для исследования примерно в середине 90-ых. Сегодня можно дать такое определение: Рекомендательные системы - это программы, которые помогают пользователю принять решение, стараясь найти из общего множества товаров, набор товаров наиболее схожий с уже понравившимся ему товарами. Что считать понравившимся товаром, зависит от области, в которой будет применяться рекомендательная система, это может быть оценка фильму или переход по ссылке на страницу продукта. Поиск таких товаров имеет первостепенную роль для онлайн бизнеса. Интернет магазины и сервисы по предоставляю контента имеют в своем распоряжении тысячи наименований, но лишь малая доля будет интересна конечному пользователю, поэтому построение правильных рекомендаций гарантированно увеличивает прибыль таких сервисов. В качестве примера можно привести интернет гиганта в сфере E-Commerce - Amazon, а так же Netflix, компанию, специализирующуюся на показе фильмов и сериалов.

1.2 Типа рекомендательных систем

Существует два основных подхода для нахождения рекомендаций: *Content-based Filtering* и *Collaborative Filtering (CF)*.

1.2.1 Content-based Filtering

Данный метод использует свойства и содержание рекомендуемых объектов. В системе основанной на данном подходе выделяют 3 основных компонента:

- **Парсер**. Этот компонент нужен для выделения нужной информации из объекта и ее структурирования. В качестве объекта может выступать веб-страница, набор действий пользователя, текстовый документ и так далее. Используются различные методы *feature extraction* для выделения ключевых свойств объекта, например, текстовый документ можно представить вектором ключевых слов.

- **Конструктор пользовательского профиля**. Этот модуль получает уже структурированные данные и пытается на их основе построить профиль пользователя.

- **Фильтр**. Данный компонент отвечает за построение конкретных рекомендаций. Имея готовый профиль пользователя и некоторую метрику, позволяющую измерить сходство между двумя объектами, строится ранжированный список наиболее похожих предметов, которые и являются результатом работы всей системы.

1.2.2 Collaborative Filtering

Данный метод основывается на большом количестве собранных от пользователей отзывов, а не свойствах рекомендуемого объекта. Главной целью является поиск схожих пользователей, основываясь на оценках, которые они поставили предмету. Найдя такие группы, довольно легко построить рекомендации.

В данной работе будет исследоваться коллаборативная фильтрация, поэтому формализуем решаемую ей задачу, она так же называется *задачей о рекомендациях.*

Прежде чем ввести формальное определение задачи, обозначим множество пользователей за $\mathcal{U}$, множество предметов за $\mathcal{I}$ и множество всех оценок за $\mathcal{R}$. Так же оговорим, что никакой пользователь не оценивал один и тот же предмет дважды, тогда оценка пользователя u предмету i запишем, как r_{ui}. Можно составить матрицу, где по строкам будут пользователи, по столбцам предметы, а элемент матрицы будет из множества оценок $\mathcal{R}$. В среднем каждый пользователь оценивает не больше 10-20 предметов, поэтому в матрице будет очень много элеметов для которых оценка r_{ui} неизвестена. Задача рекомендательной системы предсказать значение r_{ui}.

1.3 Способы реализации CF

Существует два основных метода решения задачи о рекомендациях используя CF.

- Метод соседей (Neighborhood based)
- Матричное разложение

1.3.1 Метод соседей

Данный метод очень простой в реализации и основывается на поиске множества пользователей, похожих на пользователя u. Так как каждый пользователь представляет собой вектор оценок, то схожесть двух пользователей можно оценить, посчитав косинус угла между двумя векторами. После того, как были найдены n схожих пользователей, легко пресказать непроставленные оценки пользователя u, используя имеющиеся оценки найденных пользователей.

1.3.2 Матричное разложение

Поскольку матрицу оценок пользователей может быть очень большой, а заполненных ячеек в ней очень мало, можно уменьшить размер пространства путем поиска некоторого общего набора факторов f_i которые будут общими, как для пользователей, так и для предметов. Смысл матричной факторизации заключается в приближении исходной большой матрицы произведением нескольких, но меньших по размеру.

В данной работе будет рассмотрено наиболее популярное и применяемое разложение - SVD (Singular Value Decomposition).

$$M_{m,n} = U_{m,f} x K_{f,f} x I_{n,f}$$

Здесь матрица слева - это исходная матрица оценок (m = $|mathcalU|$, n = $|mathcalI|$. Матрицы справа и есть искомое разложение, рассмотрим его подробнее. Элемент матрицы U содержит веса каждого из факторов для конкретного пользователя, другими словами, элемент $u_{i,j}$ говорит насколько важен пользователя $\mathcal{U}\rangle$ фактор f_j. Тоже самое, но для предметов, содержит матрица I. Матрица K диагональная, на ее диагонали в убывающем порядке, находятся сингулярные числа матрицы M, они показыывают, насколько фактор f_i важен, в контексте рекоментальных систем, он показывает какую долю оценки занимает фактор f_i, например, жанр

фильма может иметь большое значение для пользователя, в то время как год выпуска - нет. Сами факторы f не имеют никакой интерпритации, и SVD разложение не говорит, что это за факторы. Найдя такое разложение, не только уменьшается объем памяти, затрачиваемый на работу рекомендательной системы, но и легко решается задача о рекомендациях. Перемножив эти три матрицы, получится матрица исходная матрица оценок, но она уже будет полной, а не разреженной.

В данной работе, будет использоваться именно SVD разложение.

Глава 2

2.1 Контекстная информация

Несмотря на удобство и простоту стандартных методов коллаборативной фильтрации, данная модель не содержит в себе огромное количество другой доступной информации, будем называть такую информацию *контекстной*. Контекстной информацией может быть все что угодно, дата покупки (выходной или будний день), фанр фильма, возраст пользователя и так далее. Наличие такой информации должно повысить точность рекомендаций и их полезность. Так, например, на выходных можно рекомендовать фильмы для семейного просмотра, а по будням непродолжительные сериалы.

2.2 Встраивания контекстной информации

Существует несколько общих методов встраивания контекстной информации в рекомендательную систему.

- **Contextual Pre-Filtering**. Суть данного метода заключается в фильтрации не подходящих под текущий контекст оценок, дальше задача решается стандартными алгоритмами.
- **Contextual Post-Filtering**. После нахождения списка рекомендаций стандартными методами, предметы, не подходящие под текущий контекст, удаляются.
- **Contextual Modeling**. Контекстная информация встраивается на уровне модели. При таком способе встраивания контекстной информации, стандартные методы решения задачи о рекомендациях уже не работают или же требуется их адаптация.

В данной работе будет рассмотрен последний из методов, а именно, метод встраивания контекстной информации в матрицу отзывов.

2.3 Контекстная информация в CF

В работе будет использоваться набор данных MovieLens, он содержит отзывы на фильмы, а так же ряд дополнительной информации: возраст, пол и род деятельности пользователя, жанр и год выпуска фильма (один фильм может иметь несколько жанров).

Обозначим контекстную информацию, относящуюся к пользователю за $\mathcal{C}_{\mathcal{U}}$, а к предметам за $\mathcal{C}_{\mathcal{I}}$.

$$\mathcal{C}_{\mathcal{U}} = M, F, Job_1, ..., Job_{19}$$
$$\mathcal{C}_{\mathcal{I}} = genre_1, ..., genre_{19}$$

К исходной матрице отзывов по строкам допишем информацию о предметах, а по столбцам информацию о пользователях, таким образом, матрица отзывов размера m на n примет размер $m + |C_I|$ на $n + |C_U|$. Если предмет или пользователь содержит контекстную информацию, то в соотвествующей ячейке матрицы ставится оценка 5, иначе 0.

Таким образом мы получили , которая содержит в себе всю необходимую информацию. Отличие от стандартной матрицы заключается в том, что добавленные строчки, как бы являются пользователями, которые очень любят фильмы определенного жанра. Алгоритм CF учтет таких пользователей и найдет наиболее схожих с ними. Аналогичную аналогию можно провести и для добавленных столбцов. К новой матрице можно применить уже описанный ранее метод разложения SVD.

Глава 3

3.1 Реализация на Python

Для проверки работоспособности и эффективности подхода к контекстной информации, предложенного во второй главе был написан прототип рекомендательной системы на языке Python.

3.1.1 Реализация SVD

Обычные алгоритмы поиска SVD разложения не подходят для матрицы отзывов, потому что в такой матрице значения многих элементов неизвестны, их предсказание и является задачей системы. Однако, все классические алгоритмы поиска SVD ожидают полную матрицу. Существует два способа решения данной проблемы:

- Попытаться заполнить пустые ячейки определенными значениями, например, средней оценкой для всех фильмов или нулями. Такой метод вносит неточность в результат разложения и значительно увеличивает объем затрачиевомой памяти для хранения матрицы.
- Применить метод, предложенный Симоном Фанком [?], который основан на градиентном спуске.

3.1.2 FunkSVD

Суть предложенного Фанком алгоритма заключается в использовани метода стохастического градиентного спуска, используя только известные значения матрицы отзывов. Результатом работы алгоритма будет две матрицы, а не три, как в классическом разложении. Диагональная матрица с сингулярными значениями уже будет инкапсулирована в эти две матрицы.

Для нахождения разложения, используется сохастический градиентный спуск. Каждый признак обучается по отедльности при помощи следующих правил:

$$U_{f_i} = U_{f_i} + L(2eI_{f_i} - GU_{f_i})$$
$$I_{f_i} = I_{f_i} + L(2eU_{f_i} - GI_{f_i})$$

Здесь L - это константа, отвечающая за скорость обучения, G - нормализующая константа для борьбы с переобучением, $e = r_{kj} - (U_k, I_j)$ разница между предсказаным значением оценки и настоящим. Поскольку обучение происходит на известных оценках, то и r_{kj} всегда известен.

3.2 Оценка эффективности

Существует несколько методов оценивания работы рекомендательной системы. После того, как были найдены матрицы признаков U и I, выбирается тестовый набор данных, из него

берутся оценки пользователей, которые на момент обучения были неизвестны, и система предсказывает эти оценки $\widehat{r_i} = (U_i, I_i)$. Далее высчитываются следующие метрики:

Разница между MAE и RMSE заключается в том, что RMSE намного сильнее реагирует на большие отклонения.

Были получены следующие значения ошибок:

	RMSE	MAE
SVD	0.934347	0.758373
SVD + Context	0.919411	0.753776

Видно, что рекомендательная система, которая учитывает контекстную информацию, способна лучше предсказывать оценки пользователей.

3.3 Исходный код

Далее будет представлен исходный код. Все исходники доступны на GitHub

Литература

1. Tintarev Nava, Masthoff Judith. Recommender Systems Handbook. 2011. Т. 54. С. 479–510. URL: http://www.springerlink.com/index/10.1007/978-0-387-85820-3.